

On the acceleration of the double smoothing technique for unconstrained convex optimization problems

Radu Ioan Bot^{*}

Christopher Hendrich[†]

March 3, 2013

Abstract. In this article we investigate the possibilities of accelerating the double smoothing technique when solving unconstrained nondifferentiable convex optimization problems. This approach relies on the regularization in two steps of the Fenchel dual problem associated to the problem to be solved into an optimization problem having a differentiable strongly convex objective function with Lipschitz continuous gradient. The doubly regularized dual problem is then solved via a fast gradient method. The aim of this paper is to show how do the properties of the functions in the objective of the primal problem influence the implementation of the double smoothing approach and its rate of convergence. The theoretical results are applied to linear inverse problems by making use of different regularization functionals.

Keywords. Fenchel duality, regularization, fast gradient method, image processing

AMS subject classification. 90C25, 90C46, 47A52

1 Introduction

In this paper we are developing an efficient algorithm based on the double smoothing approach for solving unconstrained nondifferentiable optimization problems of the type

$$(P) \quad \inf_{x \in \mathcal{H}} \{f(x) + g(Ax)\}, \quad (1)$$

where \mathcal{H} is a Hilbert space, $f : \mathcal{H} \rightarrow \overline{\mathbb{R}}$ and $g : \mathbb{R}^m \rightarrow \overline{\mathbb{R}}$ are proper, convex and lower semicontinuous functions and $A : \mathcal{H} \rightarrow \mathbb{R}^m$ is a linear continuous operator fulfilling the feasibility condition $A(\text{dom } f) \cap \text{dom } g \neq \emptyset$. The double smoothing technique for solving this class of optimization problems (see [8] for a fully finite-dimensional spaces version of it) assumes to efficiently solve the corresponding Fenchel dual problems and then to recover via an approximately optimal solution of the latter an approximately optimal solution of the primal. This technique, which represents a generalization of the approach developed in [10, 11] for a special class of convex constrained optimization problems, makes use of the structure of the Fenchel dual and relies on the regularization

^{*}Faculty of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany, e-mail: radu.bot@mathematik.tu-chemnitz.de. Research partially supported by DFG (German Research Foundation), project BO 2516/4-1.

[†]Faculty of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany, e-mail: christopher.hendrich@mathematik.tu-chemnitz.de.

of the latter in two steps into an optimization problem having a differentiable strongly convex objective function with Lipschitz continuous gradient. The regularized dual is then solved by a fast gradient method which gives rise to a sequence of dual variables that solve the non-regularized dual problem after $O\left(\frac{1}{\epsilon} \ln\left(\frac{1}{\epsilon}\right)\right)$ iterations, whenever f and g have bounded effective domains. In addition, the norm of the gradient of the regularized dual objective decreases by the same rate of convergence, a fact which is crucial in view of reconstructing an approximately optimal solution to (P) after $O\left(\frac{1}{\epsilon} \ln\left(\frac{1}{\epsilon}\right)\right)$ iterations (see [8]). The first aim of this paper is to show that, whenever g is a strongly convex function, one can obtain the same convergence rate, even without imposing boundedness for its effective domain. Further we show that if, additionally, f is strongly convex or g is everywhere differentiable with a Lipschitz continuous gradient, then the convergence rate becomes $O\left(\frac{1}{\sqrt{\epsilon}} \ln\left(\frac{1}{\epsilon}\right)\right)$, while, if these supplementary assumptions are simultaneous fulfilled, then a convergence rate of $O\left(\ln\left(\frac{1}{\epsilon}\right)\right)$ can be guaranteed.

The structure of the paper is the following. The forthcoming section is dedicated to some preliminaries on convex analysis and Fenchel duality. In Section 3 we employ the smoothing technique introduced in [13–15] in order to make the objective of the Fenchel dual problem of (P) to be strongly convex and differentiable with Lipschitz continuous gradient. In Section 4 we first solve the regularized dual problem via an efficient fast gradient method. Then we show how do the properties of the functions in the objective of (P) influence the implementation of the double smoothing approach and improve its rate of convergence. We also prove how an approximately optimal primal solution can be recovered from a dual iterate. Finally, in Section 5, we consider an application of the presented approach in image deblurring and solve to this end by a linear inverse problem by using two different regularization functionals.

2 Preliminaries on convex analysis and Fenchel duality

Throughout this paper $\langle \cdot, \cdot \rangle$ and $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$ denote the *inner product* and, respectively, the *norm* of the Hilbert space \mathcal{H} , which is allowed to be infinite dimensional. The *closure* of a set $C \subseteq \mathcal{H}$ is denoted by $\text{cl}(C)$, while its *indicator function* is the function $\delta_C : \mathcal{H} \rightarrow \overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ defined by $\delta_C(x) = 0$ for $x \in C$ and $\delta_C(x) = +\infty$, otherwise. For a function $f : \mathcal{H} \rightarrow \overline{\mathbb{R}}$ we denote by $\text{dom } f := \{x \in \mathcal{H} : f(x) < +\infty\}$ its *effective domain*. We call f *proper* if $\text{dom } f \neq \emptyset$ and $f(x) > -\infty$ for all $x \in \mathcal{H}$. The *conjugate function* of f is $f^* : \mathcal{H} \rightarrow \overline{\mathbb{R}}$, $f^*(p) = \sup \{\langle p, x \rangle - f(x) : x \in \mathcal{H}\}$ for all $p \in \mathcal{H}$. The *biconjugate function* of f is $f^{**} : \mathcal{H} \rightarrow \overline{\mathbb{R}}$, $f^{**}(x) = \sup \{\langle x, p \rangle - f^*(p) : p \in \mathcal{H}\}$ and, when f is proper, convex and lower semicontinuous, then, according to the Fenchel-Moreau Theorem, one has $f = f^{**}$. The *(convex) subdifferential* of the function f at $x \in \mathcal{H}$ is the set $\partial f(x) = \{p \in \mathcal{H} : f(y) - f(x) \geq \langle p, y - x \rangle \ \forall y \in \mathcal{H}\}$, if $f(x) \in \mathbb{R}$, and is taken to be the empty set, otherwise.

Further, we consider the space \mathbb{R}^m endowed with the Euclidean inner product and norm, for which we use the same notations as for the Hilbert space \mathcal{H} , since no confusion can arise. By $\mathbf{1}^m$ we denote the vector in \mathbb{R}^m with all entries equal to 1. For a subset C of \mathbb{R}^m we denote by $\text{ri}(C)$ its *relative interior*, i.e. the interior of the set C relative to its affine hull. For a linear continuous operator $A : \mathcal{H} \rightarrow \mathbb{R}^m$ the operator $A^* : \mathbb{R}^m \rightarrow \mathcal{H}$,

defined by $\langle A^*y, x \rangle = \langle y, Ax \rangle$ for all $x \in \mathcal{H}$ and all $y \in \mathbb{R}^m$, is its so-called *adjoint operator*. By $\text{id} : \mathbb{R}^m \rightarrow \mathbb{R}^m, \text{id}(x) = x$, for all $x \in \mathbb{R}^m$ we denote the *identity mapping* on \mathbb{R}^m .

For a nonempty, convex and closed set $C \subseteq \mathcal{H}$ we consider the *projection operator* $\mathcal{P}_C : \mathcal{H} \rightarrow C$ defined as $x \mapsto \arg \min_{z \in C} \|x - z\|$. Having two functions $f, g : \mathcal{H} \rightarrow \overline{\mathbb{R}}$, their *infimal convolution* is defined by $f \square g : \mathcal{H} \rightarrow \overline{\mathbb{R}}$, $(f \square g)(x) = \inf_{y \in \mathcal{H}} \{f(y) + g(x - y)\}$ for all $x \in \mathcal{H}$. The *Moreau envelope* ${}^\gamma f : \mathcal{H} \rightarrow \overline{\mathbb{R}}$ of the function $f : \mathcal{H} \rightarrow \overline{\mathbb{R}}$ of parameter $\gamma > 0$ is defined as the infimal convolution

$${}^\gamma f(x) := f \square \left(\frac{1}{2\gamma} \|\cdot\|^2 \right) (x) = \inf_{y \in \mathcal{H}} \left\{ f(y) + \frac{1}{2\gamma} \|x - y\|^2 \right\} \quad \forall x \in \mathcal{H}.$$

For $\rho > 0$ we say that the function $f : \mathcal{H} \rightarrow \overline{\mathbb{R}}$ is ρ -strongly convex, if for all $x, y \in \mathcal{H}$ and all $\lambda \in (0, 1)$ it holds

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \frac{\rho}{2} \lambda(1 - \lambda) \|x - y\|^2.$$

Notice that this is equivalent to saying that $x \mapsto f(x) - \frac{\rho}{2} \|x\|^2$ is convex.

For the optimization problem (P) we consider the following *standing assumptions*: $f : \mathcal{H} \rightarrow \overline{\mathbb{R}}$ is a proper, convex and lower semicontinuous function with a bounded effective domain, $g : \mathbb{R}^m \rightarrow \overline{\mathbb{R}}$ is proper, μ -strongly convex ($\mu > 0$) and lower semicontinuous function and $A : \mathcal{H} \rightarrow \mathbb{R}^m$ is a linear operator fulfilling $A(\text{dom } f) \cap \text{dom } g \neq \emptyset$.

Remark 1. Different to the investigations made in [8] in a fully finite-dimensional setting, we strengthen here the convexity assumptions on g (there g was asked to be only proper, convex and lower semicontinuous), but allow in counterpart $\text{dom } g$ to be unbounded.

The Fenchel dual problem to (P) (see, for instance, [5, 6]) reads

$$(D) \quad \sup_{p \in \mathbb{R}^m} \{-f^*(A^*p) - g^*(-p)\}. \quad (2)$$

We denote the optimal objective values of the optimization problems (P) and (D) by $v(P)$ and $v(D)$, respectively.

The conjugate functions of f and g can be written as

$$f^*(q) = \sup_{x \in \text{dom } f} \{\langle q, x \rangle - f(x)\} = - \inf_{x \in \text{dom } f} \{\langle -q, x \rangle + f(x)\} \quad \forall q \in \mathcal{H}$$

and

$$g^*(p) = \sup_{x \in \text{dom } g} \{\langle p, x \rangle - g(x)\} = - \inf_{x \in \text{dom } g} \{\langle -p, x \rangle + g(x)\} \quad \forall p \in \mathbb{R}^m,$$

respectively. According to [1, Theorem 11.9] and [4, Lemma 2.33], the optimization problems arising in the formulation of both $f^*(q)$ for all $q \in \mathcal{H}$ and $g^*(p)$ for all $p \in \mathbb{R}^m$ are solvable, fact which implies that $\text{dom } f^* = \mathcal{H}$ and $\text{dom } g^* = \mathbb{R}^m$, respectively.

By writing the dual problem (D) equivalently as the infimum optimization problem

$$\inf_{p \in \mathbb{R}^m} \{f^*(A^*p) + g^*(-p)\},$$

one can easily see that the Fenchel dual problem of the latter is

$$\sup_{x \in \mathcal{H}} \{-f^{**}(x) - g^{**}(Ax)\},$$

which, by the Fenchel-Moreau Theorem, is nothing else than

$$\sup_{x \in \mathcal{H}} \{-f(x) - g(Ax)\}.$$

In order to guarantee strong duality for this primal-dual pair it is sufficient to ensure that (see, for instance, [5, Theorem 2.1]) $0 \in \text{ri}(A^*(\text{dom } g^*) + \text{dom } f^*)$. As f^* has full domain, this regularity condition is automatically fulfilled, which means that $v(D) = v(P)$ and the primal optimization problem (P) has an optimal solution. Due to the fact that f and g are proper and $A(\text{dom } f) \cap \text{dom } g \neq \emptyset$, this further implies $v(D) = v(P) \in \mathbb{R}$. Later we will assume that the dual problem (D) has an optimal solution, too, and that an upper bound of its norm is known.

Denote by $\theta : \mathbb{R}^m \rightarrow \mathbb{R}$, $\theta(p) = f^*(A^*p) + g^*(-p)$, the objective function of (D) . Hence, the dual can be equivalently written as

$$(D) \quad - \inf_{p \in \mathbb{R}^m} \theta(p). \quad (3)$$

The assumptions made on g yields that $p \mapsto g^*(-p)$ is differentiable and has a Lipschitz continuous gradient (see Subsection 3.1 for details). However, since in general one can not guarantee the smoothness of $p \mapsto f^*(A^*p)$, the dual problem (D) is a nondifferentiable convex optimization problem. Our goal is to solve this problem efficiently and to obtain from here an optimal solution to (P) . As in [8], we are overcoming the non-satisfactory complexity of subgradient-schemes, i.e. $O\left(\frac{1}{\epsilon^2}\right)$, by making use of smoothing techniques introduced in [13–15]. More precisely, we regularize first the objective function of $f^*(A^*p)$ by a quadratic term in order to obtain a smooth approximation of $p \mapsto f^*(A^*p)$. Then we apply a second regularization to the new dual objective and minimize the regularized problem via an appropriate fast gradient scheme (see [8]). This will allow us to solve both optimization problems (D) and (P) approximately in $O\left(\frac{1}{\epsilon} \ln\left(\frac{1}{\epsilon}\right)\right)$ iterations. More than that, we will show that this rate of convergence can be improved when strengthening the assumptions imposed on f and g .

3 The double smoothing approach

3.1 First smoothing

For a real number $\rho > 0$ the function $p \mapsto f^*(A^*p) = \sup_{x \in \mathcal{H}} \{\langle A^*p, x \rangle - f(x)\}$ can be approximated by

$$f_\rho^*(A^*p) = \sup_{x \in \mathcal{H}} \left\{ \langle A^*p, x \rangle - f(x) - \frac{\rho}{2} \|x\|^2 \right\}. \quad (4)$$

For each $p \in \mathbb{R}^m$ the maximization problem which occurs in the formulation of $f_\rho^*(A^*p)$ has a unique solution (see, for instance, [1, Proposition 11.14]), fact which implies that $f_\rho^*(A^*p) \in \mathbb{R}$.

For all $p \in \mathbb{R}^m$ one can express the above regularization of the conjugate by means the Moreau envelope of f as follows

$$\begin{aligned} -f_\rho^*(A^*p) &= -\sup_{x \in \mathcal{H}} \left\{ \langle A^*p, x \rangle - f(x) - \frac{\rho}{2} \|x\|^2 \right\} \\ &= \inf_{x \in \mathcal{H}} \left\{ -\langle A^*p, x \rangle + f(x) + \frac{\rho}{2} \|x\|^2 \right\} \\ &= \inf_{x \in \mathcal{H}} \left\{ f(x) + \frac{\rho}{2} \left\| \frac{A^*p}{\rho} - x \right\|^2 \right\} - \frac{\|A^*p\|^2}{2\rho} = \frac{1}{\rho} f \left(\frac{A^*p}{\rho} \right) - \frac{\|A^*p\|^2}{2\rho}. \end{aligned}$$

Consequently, one can transfer the differentiability properties of the Moreau envelope (see [1, Proposition 12.29]) to $p \mapsto -(f_\rho^* \circ A^*)(p)$. For all $p \in \mathbb{R}^m$ we have

$$-\nabla(f_\rho^* \circ A^*)(p) = \frac{A}{\rho} \nabla \frac{1}{\rho} f \left(\frac{A^*p}{\rho} \right) - \frac{AA^*p}{\rho} = \frac{A}{\rho} \left(\rho \left(\frac{A^*p}{\rho} - x_{f,p} \right) \right) - \frac{AA^*p}{\rho} = -Ax_{f,p},$$

thus

$$\nabla(f_\rho^* \circ A^*)(p) = Ax_{f,p},$$

where $x_{f,p} \in \mathcal{H}$ is the *proximal point* of parameter $\frac{1}{\rho}$ of f at $\frac{A^*p}{\rho}$, namely the unique element in \mathcal{H} fulfilling (see [1, Proposition 12.29])

$$\frac{1}{\rho} f \left(\frac{A^*p}{\rho} \right) = f(x_{f,p}) + \frac{\rho}{2} \left\| \frac{A^*p}{\rho} - x_{f,p} \right\|^2.$$

By taking into account the nonexpansiveness of the proximal point mapping (see [1, Proposition 12.27]), for $p, q \in \mathbb{R}^m$ it holds

$$\begin{aligned} \left\| \nabla(f_\rho^* \circ A^*)(p) - \nabla(f_\rho^* \circ A^*)(q) \right\| &= \|Ax_{f,p} - Ax_{f,q}\| \leq \|A\| \|x_{f,p} - x_{f,q}\| \\ &\leq \|A\| \left\| \frac{A^*p}{\rho} - \frac{A^*q}{\rho} \right\| \leq \frac{\|A\|^2}{\rho} \|p - q\|, \end{aligned}$$

thus $\frac{\|A\|^2}{\rho}$ is the Lipschitz constant of $p \mapsto \nabla(f_\rho^* \circ A^*)(p)$.

Coming now to the function $p \mapsto g^*(-p) = (g^* \circ -\text{id})(p)$, let us notice first that, since g is proper, μ -strongly convex and lower semicontinuous, g^* is differentiable and ∇g^* is Lipschitz continuous with Lipschitz constant $\frac{1}{\mu}$. Thus $(g^* \circ -\text{id})$ is Fréchet differentiable, too, and its gradient is Lipschitz continuous with Lipschitz constant $\frac{1}{\mu}$. By denoting

$$x_{g,p} := \nabla g^*(-p) = -\nabla(g^* \circ -\text{id})(p),$$

one has that $-p \in \partial g(x_{g,p})$ or, equivalently, $0 \in \partial(\langle p, \cdot \rangle + g)(x_{g,p})$, which means that $x_{g,p}$ is the unique optimal solution (see [4, Lemma 2.33]) of the optimization problem

$$\inf_{x \in \mathbb{R}^m} \{ \langle p, x \rangle + g(x) \}.$$

Remark 2. If f is ρ -strongly convex, for $\rho > 0$, then there is no need to apply the first regularization for $p \mapsto f^*(A^*p)$, as this function is already Fréchet differentiable with a Lipschitz continuous gradient having a Lipschitz constant given by $\frac{\|A\|^2}{\rho}$. Indeed, the ρ -strong convexity of f implies that f^* is Fréchet differentiable with Lipschitz continuous gradient having a Lipschitz constant given by $\frac{1}{\rho}$ (see [1, Theorem 18.15]). Hence, for all $p, q \in \mathbb{R}^m$, we have

$$\begin{aligned} \|\nabla(f^* \circ A^*)(p) - \nabla(f^* \circ A^*)(q)\| &= \|A\nabla f^*(A^*p) - A\nabla f^*(A^*q)\| \\ &\leq \frac{\|A\|}{\rho} \|A^*p - A^*q\| \leq \frac{\|A\|^2}{\rho} \|p - q\|. \end{aligned}$$

By denoting

$$x_{f,p} := \nabla f^*(A^*p),$$

one has that $0 \in \partial(f - \langle A^*p, \cdot \rangle)(x_{f,p})$, which means that $x_{f,p}$ is the unique optimal solution (see [4, Lemma 2.33]) of the optimization problem

$$\inf_{x \in \mathcal{H}} \{f(x) - \langle A^*p, x \rangle\}.$$

By denoting $D_f := \sup \left\{ \frac{\|x\|^2}{2} : x \in \text{dom } f \right\} \in \mathbb{R}$ we can relate $f^* \circ A^*$ and its smooth approximation $f_\rho^* \circ A^*$ as follows.

Proposition 3. *For all $p \in \mathbb{R}^m$ it holds*

$$f_\rho^*(A^*p) \leq f^*(A^*p) \leq f_\rho^*(A^*p) + \rho D_f.$$

Proof. For $p \in \mathbb{R}^m$ one has

$$\begin{aligned} f_\rho^*(A^*p) &= \langle A^*p, x_{f,p} \rangle - f(x_{f,p}) - \frac{\rho}{2} \|x_{f,p}\|^2 \leq \langle A^*p, x_{f,p} \rangle - f(x_{f,p}) \leq f^*(A^*p) \\ &\leq \sup_{x \in \text{dom } f} \left\{ \langle A^*p, x \rangle - f(x) - \frac{\rho}{2} \|x\|^2 \right\} + \sup_{x \in \text{dom } f} \left\{ \frac{\rho}{2} \|x\|^2 \right\} \\ &= f_\rho^*(A^*p) + \rho D_f. \end{aligned}$$

□

For $\rho > 0$ let $\theta_\rho : \mathbb{R}^m \rightarrow \mathbb{R}$ be defined by $\theta_\rho(p) = f_\rho^*(A^*p) + g^*(-p)$. The function θ_ρ is differentiable with a Lipschitz continuous gradient

$$\nabla \theta_\rho(p) = \nabla(f_\rho^* \circ A^*)(p) + \nabla(g^* \circ -\text{id})(p) = Ax_{f,p} - x_{g,p} \quad \forall p \in \mathbb{R}^m,$$

having as Lipschitz constant $L(\rho) := \frac{\|A\|^2}{\rho} + \frac{1}{\mu}$.

In consideration of Proposition 3 we get

$$\theta_\rho(p) \leq \theta(p) \leq \theta_\rho(p) + \rho D_f \quad \forall p \in \mathbb{R}^m. \quad (5)$$

In order to reconstruct an approximately optimal solution to the primal optimization problem (P) it is not sufficient to ensure the convergence of $\theta(\cdot)$ to $-v(D)$, but we also need good convergence properties for the decrease of $\|\nabla \theta_\rho(\cdot)\|$ (cf. [8, 10]).

3.2 Second smoothing

In the following, a second regularization is applied to θ_ρ , as done in [8, 10, 11], in order to make it strongly convex, fact which will allow us to use a fast gradient scheme with a good convergence rate for the decrease of $\|\nabla\theta_\rho(\cdot)\|$. Therefore, adding the strongly convex function $\frac{\kappa}{2}\|\cdot\|^2$ to θ_ρ , for some positive real number κ , gives rise to the following regularization of the objective function

$$\theta_{\rho,\kappa} : \mathbb{R}^m \rightarrow \mathbb{R}, \quad \theta_{\rho,\kappa}(p) := \theta_\rho(p) + \frac{\kappa}{2} \|p\|^2 = f_\rho^*(A^*p) + g^*(-p) + \frac{\kappa}{2} \|p\|^2,$$

which is obviously κ -strongly convex. We further deal with the optimization problem

$$\inf_{p \in \mathbb{R}^m} \theta_{\rho,\kappa}(p). \quad (6)$$

By taking into account [4, Lemma 2.33], the optimization problem (6) has a unique optimal solution, while the function $\theta_{\rho,\kappa}$ is differentiable and for all $p \in \mathbb{R}^m$ it holds

$$\nabla\theta_{\rho,\kappa}(p) = \nabla \left(\theta_\rho(\cdot) + \frac{\kappa}{2} \|\cdot\|^2 \right) (p) = Ax_{f,p} - x_{g,p} + \kappa p.$$

This gradient is Lipschitz continuous with constant $L(\rho, \kappa) := \frac{\|A\|^2}{\rho} + \frac{1}{\mu} + \kappa$.

Remark 4. If θ_ρ is κ -strongly convex, then there is no need to apply the second regularization, as this function is already endowed with the properties of $\theta_{\rho,\kappa}$.

4 Solving the doubly regularized dual problem

4.1 A fast gradient method

In the forthcoming sections we denote by p_{DS}^* the unique optimal solution of the optimization problem (6) and by $\theta_{\rho,\kappa}^* := \theta_{\rho,\kappa}(p_{DS}^*)$ its optimal objective value. Further, we denote by $p^* \in \mathbb{R}^m$ an optimal solution to the dual optimization problem (D) and we assume that the upper bound

$$\|p^*\| \leq R \quad (7)$$

is available for some nonzero $R \in \mathbb{R}_+$.

Furthermore, we make use of the following fast gradient method (see [12, Algorithm 2.2.11])

$$\begin{aligned} \text{Init.:} \quad & \text{Set } w_0 = p_0 := 0 \in \mathbb{R}^m \\ \text{For } k \geq 0 : \quad & \text{Set } p_{k+1} := w_k - \frac{1}{L(\rho, \kappa)} \nabla\theta_{\rho,\kappa}(w_k). \\ & \text{Set } w_{k+1} := p_{k+1} + \frac{\sqrt{L(\rho, \kappa)} - \sqrt{\kappa}}{\sqrt{L(\rho, \kappa)} + \sqrt{\kappa}} (p_{k+1} - p_k) \end{aligned} \quad (8)$$

for minimizing the optimization problem (6), which has a strongly convex and differentiable optimization function with a Lipschitz continuous gradient. By taking into

account [12, Theorem 2.2.3] we obtain a sequence $(p_k)_{k \geq 0} \subseteq \mathbb{R}^m$ satisfying

$$\begin{aligned} \theta_{\rho, \kappa}(p_k) - \theta_{\rho, \kappa}^* &\leq \left(\theta_{\rho, \kappa}(p_0) - \theta_{\rho, \kappa}^* + \frac{\kappa}{2} \|p_0 - p_{DS}^*\|^2 \right) \left(1 - \sqrt{\frac{\kappa}{L(\rho, \kappa)}} \right)^k \\ &\leq (\theta_{\rho, \kappa}(p_0) - \theta_{\rho, \kappa}^* + \frac{\kappa}{2} \|p_0 - p_{DS}^*\|^2) e^{-k \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \end{aligned} \quad (9)$$

$$\leq 2(\theta_{\rho, \kappa}(p_0) - \theta_{\rho, \kappa}^*) e^{-k \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \quad \forall k \geq 0, \quad (10)$$

while the last inequality is a consequence of [12, Theorem 2.1.8]. Since p_{DS}^* solves (6), we have $\nabla \theta_{\rho, \kappa}(p_{DS}^*) = 0$ and therefore [12, Theorem 2.1.5] yields

$$\frac{1}{2L(\rho, \kappa)} \|\nabla \theta_{\rho, \kappa}(p_k)\|^2 \leq \theta_{\rho, \kappa}(p_k) - \theta_{\rho, \kappa}^* \stackrel{(10)}{\leq} 2(\theta_{\rho, \kappa}(p_0) - \theta_{\rho, \kappa}^*) e^{-k \sqrt{\frac{\kappa}{L(\rho, \kappa)}}},$$

which implies

$$\|\nabla \theta_{\rho, \kappa}(p_k)\|^2 \leq 4L(\rho, \kappa)(\theta_{\rho, \kappa}(p_0) - \theta_{\rho, \kappa}^*) e^{-k \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \quad \forall k \geq 0. \quad (11)$$

Due to the κ -strong convexity of $\theta_{\rho, \kappa}$, [12, Theorem 2.1.8] states

$$\frac{\kappa}{2} \|p_k - p_{DS}^*\|^2 \leq \theta_{\rho, \kappa}(p_k) - \theta_{\rho, \kappa}^* \stackrel{(10)}{\leq} 2(\theta_{\rho, \kappa}(p_0) - \theta_{\rho, \kappa}^*) e^{-k \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \quad \forall k \geq 0. \quad (12)$$

Using this inequality it follows that (see also [10, 11])

$$\|p_k - p_{DS}^*\|^2 \leq \min \left\{ \|p_0 - p_{DS}^*\|^2, \frac{4}{\kappa} (\theta_{\rho, \kappa}(p_0) - \theta_{\rho, \kappa}^*) e^{-k \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \right\} \quad \forall k \geq 0. \quad (13)$$

We first prove that the rates of convergence for the decrease of $\theta(p_k) - \theta(p^*)$ and $\|\nabla \theta_{\rho}(p_k)\|$ coincide, being equal to $O\left(\frac{1}{\epsilon} \ln\left(\frac{1}{\epsilon}\right)\right)$, and that they can be improved when f and/or g fulfill additional assumptions. We also show how ϵ -optimal solutions to the primal problem (P) can be recovered from the sequence of dual variables $(p_k)_{k \geq 0}$.

4.2 Convergence of $\theta(p_k)$ to $\theta(p^*)$

Since the algorithm starts with $p_0 = 0$, we have $\theta_{\rho, \kappa}(0) = f_{\rho}^*(0) + g^*(0) + \frac{\kappa}{2} \|0\|^2 = \theta_{\rho}(0)$, while

$$\theta_{\rho, \kappa}(p_{DS}^*) = \theta_{\rho}(p_{DS}^*) + \frac{\kappa}{2} \|p_{DS}^*\|^2. \quad (14)$$

Making use of these two relations we obtain

$$\frac{\kappa}{2} \|p_{DS}^*\|^2 \stackrel{(12)}{\leq} \theta_{\rho, \kappa}(0) - \theta_{\rho, \kappa}(p_{DS}^*) = \theta_{\rho}(0) - \theta_{\rho}(p_{DS}^*) - \frac{\kappa}{2} \|p_{DS}^*\|^2,$$

which further implies that

$$\|p_{DS}^*\|^2 \leq \frac{1}{\kappa} (\theta_{\rho}(0) - \theta_{\rho}(p_{DS}^*)). \quad (15)$$

Additionally, in all iterations $k \geq 0$, we have

$$\begin{aligned}
\|p_k - p_{DS}^*\|^2 &\stackrel{(12)}{\leq} \frac{2}{\kappa} (\theta_{\rho, \kappa}(p_k) - \theta_{\rho, \kappa}(p_{DS}^*)) \\
&\stackrel{(9)}{\leq} \frac{2}{\kappa} \left(\theta_{\rho, \kappa}(0) - \theta_{\rho, \kappa}(p_{DS}^*) + \frac{\kappa}{2} \|0 - p_{DS}^*\|^2 \right) e^{-k \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \\
&\stackrel{(14)}{=} \frac{2}{\kappa} (\theta_{\rho}(0) - \theta_{\rho}(p_{DS}^*)) e^{-k \sqrt{\frac{\kappa}{L(\rho, \kappa)}}}
\end{aligned} \tag{16}$$

and

$$\begin{aligned}
\theta_{\rho}(p_k) - \theta_{\rho}(p_{DS}^*) &\stackrel{(9)}{\leq} \left(\theta_{\rho, \kappa}(0) - \theta_{\rho, \kappa}(p_{DS}^*) + \frac{\kappa}{2} \|0 - p_{DS}^*\|^2 \right) e^{-k \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \\
&\quad + \frac{\kappa}{2} (\|p_{DS}^*\|^2 - \|p_k\|^2) \\
&\stackrel{(14)}{=} (\theta_{\rho}(0) - \theta_{\rho}(p_{DS}^*)) e^{-k \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} + \frac{\kappa}{2} (\|p_{DS}^*\|^2 - \|p_k\|^2).
\end{aligned} \tag{17}$$

The estimation

$$\begin{aligned}
\|p_{DS}^*\|^2 - \|p_k\|^2 &= (\|p_{DS}^*\| - \|p_k\|) (\|p_{DS}^*\| + \|p_k\|) \\
&\leq \|p_{DS}^* - p_k\| (\|p_{DS}^*\| + \|p_k\|) \\
&\leq \|p_{DS}^* - p_k\| (2\|p_{DS}^*\| + \|p_k - p_{DS}^*\|) \\
&\stackrel{(13)}{\leq} 3\|p_{DS}^* - p_k\| \|p_{DS}^*\| \\
&\stackrel{(16)}{\leq} 3\|p_{DS}^*\| \sqrt{\frac{2}{\kappa} (\theta_{\rho}(0) - \theta_{\rho}(p_{DS}^*))} e^{-\frac{k}{2} \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \\
&\stackrel{(15)}{\leq} \frac{3\sqrt{2}}{\kappa} (\theta_{\rho}(0) - \theta_{\rho}(p_{DS}^*)) e^{-\frac{k}{2} \sqrt{\frac{\kappa}{L(\rho, \kappa)}}}
\end{aligned}$$

can now be inserted into (17) and this leads to

$$\begin{aligned}
\theta_{\rho}(p_k) - \theta_{\rho}(p_{DS}^*) &\leq (\theta_{\rho}(0) - \theta_{\rho}(p_{DS}^*)) \left(e^{-k \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} + \frac{3}{\sqrt{2}} e^{-\frac{k}{2} \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \right) \\
&\leq \frac{25}{8} (\theta_{\rho}(0) - \theta_{\rho}(p_{DS}^*)) e^{-\frac{k}{2} \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \quad \forall k \geq 0.
\end{aligned} \tag{18}$$

Further, we have $\theta_{\rho}(0) \stackrel{(5)}{\leq} \theta(0)$, $\theta_{\rho}(p_{DS}^*) \stackrel{(5)}{\geq} \theta(p_{DS}^*) - \rho D_f \geq \theta(p^*) - \rho D_f$ and, from here,

$$\theta_{\rho}(0) - \theta_{\rho}(p_{DS}^*) \leq \theta(0) - \theta(p^*) + \rho D_f. \tag{19}$$

Since $\theta_{\rho}(p_{DS}^*) \leq \theta_{\rho}(p^*) + \frac{\kappa}{2} \|p_{DS}^*\|^2 \leq \theta_{\rho}(p^*) + \frac{\kappa}{2} \|p^*\|^2$, we obtain that

$$\theta_{\rho}(p_{DS}^*) \leq \theta_{\rho}(p^*) + \frac{\kappa}{2} \|p^*\|^2 \stackrel{(5)}{\leq} \theta(p^*) + \frac{\kappa}{2} \|p^*\|^2$$

and, therefore,

$$\theta_{\rho}(p_k) - \theta_{\rho}(p_{DS}^*) \stackrel{(5)}{\geq} \theta(p_k) - \rho D_f - \theta(p^*) - \frac{\kappa}{2} \|p^*\|^2 \quad \forall k \geq 0. \tag{20}$$

In conclusion we obtain for all $k \geq 0$

$$\begin{aligned}
\theta(p_k) - \theta(p^*) &\stackrel{(20)}{\leq} \rho D_f + \frac{\kappa}{2} \|p^*\|^2 + \theta_\rho(p_k) - \theta_\rho(p_{DS}^*) \\
&\stackrel{(7),(18)}{\leq} \rho D_f + \frac{\kappa}{2} R^2 + \frac{25}{8} (\theta_\rho(0) - \theta_\rho(p_{DS}^*)) e^{-\frac{k}{2} \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \\
&\stackrel{(19)}{\leq} \rho D_f + \frac{\kappa}{2} R^2 + \frac{25}{8} (\theta(0) - \theta(p^*) + \rho D_f) e^{-\frac{k}{2} \sqrt{\frac{\kappa}{L(\rho, \kappa)}}}. \quad (21)
\end{aligned}$$

Next we fix $\epsilon > 0$. In order to get $\theta(p_k) - \theta(p^*) \leq \epsilon$ after a certain amount of iterations k , we force all three terms in (21) to be less than or equal to $\frac{\epsilon}{3}$. To this end we choose first

$$\rho := \rho(\epsilon) = \frac{\epsilon}{3D_f} \text{ and } \kappa := \kappa(\epsilon) = \frac{2\epsilon}{3R^2}. \quad (22)$$

With these new parameters we can simplify (21) to

$$\theta(p_k) - \theta(p^*) \leq \frac{2\epsilon}{3} + \frac{25}{8} \left(\theta(0) - \theta(p^*) + \frac{\epsilon}{3} \right) e^{-\frac{k}{2} \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \quad \forall k \geq 0,$$

thus, the second term in the expression on the right-hand side of the above estimate determines the number of iterations needed to obtain ϵ -accuracy for the dual objective function θ . Indeed, we have

$$\begin{aligned}
\frac{\epsilon}{3} &\geq \frac{25}{8} \left(\theta(0) - \theta(p^*) + \frac{\epsilon}{3} \right) e^{-\frac{k}{2} \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \\
\Leftrightarrow e^{\frac{k}{2} \sqrt{\frac{\kappa}{L(\rho, \kappa)}}} &\geq \frac{3}{\epsilon} \cdot \frac{25}{8} \left(\theta(0) - \theta(p^*) + \frac{\epsilon}{3} \right) \\
\Leftrightarrow \frac{k}{2} \sqrt{\frac{\kappa}{L(\rho, \kappa)}} &\geq \ln \left(\frac{75 (\theta(0) - \theta(p^*) + \frac{\epsilon}{3})}{8\epsilon} \right) \\
\Leftrightarrow k &\geq 2 \sqrt{\frac{L(\rho, \kappa)}{\kappa}} \ln \left(\frac{75 (\theta(0) - \theta(p^*) + \frac{\epsilon}{3})}{8\epsilon} \right). \quad (23)
\end{aligned}$$

Noticing that

$$\begin{aligned}
\frac{L(\rho, \kappa)}{\kappa} &= \frac{\|A\|^2}{\rho\kappa} + \frac{1}{\mu\kappa} + 1 \stackrel{(22)}{=} \frac{9\|A\|^2 D_f R^2}{2\epsilon^2} + \frac{3R^2}{2\mu\epsilon} + 1 \\
&= \frac{1}{\epsilon^2} \left(\frac{9\|A\|^2 D_f R^2}{2} + \frac{3R^2\epsilon}{2\mu} + \epsilon^2 \right),
\end{aligned}$$

in order to obtain an approximately optimal solution to (D) , we need $k = O\left(\frac{1}{\epsilon} \ln\left(\frac{1}{\epsilon}\right)\right)$ iterations.

4.3 Convergence of $\|\nabla\theta_\rho(p_k)\|$ to 0

Guaranteeing ϵ -optimality for the objective value of the dual is not sufficient for solving the primal optimization problem with a good convergence rate, as we need at least the

same convergence rate for the decrease of $\|\nabla\theta_\rho(p_k)\| = \|Ax_{f,p_k} - x_{g,p_k}\|$ to 0. Within this section we show that this desiderate is attained (see also [10, 11]). Since

$$\|p_k\| = \|p_k - p_{DS}^* + p_{DS}^*\| \leq \|p_k - p_{DS}^*\| + \|p_{DS}^*\| \stackrel{(13)}{\leq} 2\|p_{DS}^*\|,$$

we conclude that

$$\begin{aligned} \|\nabla\theta_\rho(p_k)\| &\leq \|\nabla\theta_{\rho,\kappa}(p_k)\| + \|\kappa p_k\| \\ &\leq \|\nabla\theta_{\rho,\kappa}(p_k)\| + 2\kappa\|p_{DS}^*\| \quad \forall k \geq 0. \end{aligned} \quad (24)$$

We further have

$$\begin{aligned} \|\nabla\theta_{\rho,\kappa}(p_k)\|^2 &\stackrel{(11)}{\leq} 4L(\rho, \kappa)(\theta_{\rho,\kappa}(0) - \theta_{\rho,\kappa}(p_{DS}^*)) e^{-k\sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \\ &\stackrel{(14)}{\leq} 4L(\rho, \kappa)(\theta_\rho(0) - \theta_\rho(p_{DS}^*)) e^{-k\sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \\ &\stackrel{(19)}{\leq} 4L(\rho, \kappa) \left(\theta(0) - \theta(p^*) + \frac{\epsilon}{3} \right) e^{-k\sqrt{\frac{\kappa}{L(\rho, \kappa)}}}, \end{aligned}$$

which yields

$$\|\nabla\theta_{\rho,\kappa}(p_k)\| \leq 2\sqrt{L(\rho, \kappa) \left(\theta(0) - \theta(p^*) + \frac{\epsilon}{3} \right)} e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \quad \forall k \geq 0. \quad (25)$$

In order to give an upper bound for the second term in (24), we notice that

$$\begin{aligned} \theta(p^*) + \frac{\kappa}{2}\|p^*\|^2 &\stackrel{(5)}{\geq} \theta_\rho(p^*) + \frac{\kappa}{2}\|p^*\|^2 \geq \theta_\rho(p_{DS}^*) + \frac{\kappa}{2}\|p_{DS}^*\|^2 \\ &\stackrel{(5)}{\geq} \theta(p_{DS}^*) - \rho D_f + \frac{\kappa}{2}\|p_{DS}^*\|^2 \\ &\geq \theta(p^*) - \rho D_f + \frac{\kappa}{2}\|p_{DS}^*\|^2, \end{aligned}$$

which is equivalent to $\frac{\kappa}{2}\|p_{DS}^*\|^2 \leq \frac{\kappa}{2}\|p^*\|^2 + \rho D_f$, i. e. $\|p_{DS}^*\|^2 \leq \|p^*\|^2 + \frac{2\rho}{\kappa} D_f$. Hence,

$$\|p_{DS}^*\| \leq \sqrt{\|p^*\|^2 + \frac{2\rho}{\kappa} D_f} \stackrel{(22)}{=} \sqrt{\|p^*\|^2 + \frac{2\epsilon}{3\kappa}} \stackrel{(22)}{=} \sqrt{\|p^*\|^2 + R^2} \stackrel{(7)}{\leq} \sqrt{2}R, \quad (26)$$

which, combined with (24) and (25), provides

$$\begin{aligned} \|\nabla\theta_\rho(p_k)\| &\leq 2\sqrt{L(\rho, \kappa) \left(\theta(0) - \theta(p^*) + \frac{\epsilon}{3} \right)} e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho, \kappa)}}} + 2\sqrt{2}\kappa R \\ &= 2\sqrt{L(\rho, \kappa) \left(\theta(0) - \theta(p^*) + \frac{\epsilon}{3} \right)} e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho, \kappa)}}} + \frac{4\sqrt{2}\epsilon}{3R} \quad \forall k \geq 0. \end{aligned} \quad (27)$$

For $\epsilon > 0$ fixed, the first term in (27) decreases by the iteration counter k , and, by taking into account (22), we can ensure

$$\theta(p_k) - \theta(p^*) \leq \epsilon \quad \text{and} \quad \|\nabla\theta_\rho(p_k)\| \leq \frac{2\epsilon}{R} \quad (28)$$

in $k = O\left(\frac{1}{\epsilon} \ln\left(\frac{1}{\epsilon}\right)\right)$ iterations.

4.4 Improved convergence rates

In this subsection we investigate how additionally assumptions on the functions f and/or g influence the implementation of the double smoothing approach and its rate of convergence.

4.4.1 The case f is strongly convex

Assuming additionally to the standing assumptions that the function $f : \mathcal{H} \rightarrow \overline{\mathbb{R}}$ is ρ -strongly convex, for $\rho > 0$, the first smoothing, as done in Subsection 3.1, can be omitted and the fast gradient method (8) can be applied to the function $\theta_\kappa : \mathbb{R}^m \rightarrow \mathbb{R}$, $\theta_\kappa := f^*(A^*p) + g^*(-p) + \frac{\kappa}{2} \|p\|^2$, with $\kappa > 0$, which is κ -strongly convex and differentiable with Lipschitz continuous gradient. In the light of Remark 2 the Lipschitz constant of $\nabla\theta_\kappa$ is $L(\kappa) := \frac{\|A\|^2}{\rho} + \frac{1}{\mu} + \kappa$.

Similar to the calculations made in Section 4.2 we obtain for all $k \geq 0$

$$\theta(p_k) - \theta(p^*) \leq \frac{\kappa}{2} R^2 + \frac{25}{8} (\theta(0) - \theta(p^*)) e^{-\frac{k}{2} \sqrt{\frac{\kappa}{L(\kappa)}}}.$$

Hence, when $\varepsilon > 0$, in order to guarantee ε -accuracy for the dual objective function we can force both terms in the above estimate to be less than or equal to $\frac{\varepsilon}{2}$. Thus, by taking

$$\kappa := \kappa(\varepsilon) = \frac{\varepsilon}{R^2},$$

this time we will need to this end, in contrast to (23),

$$k \geq 2 \sqrt{\frac{L(\kappa)}{\kappa}} \ln \left(\frac{25 (\theta(0) - \theta(p^*))}{4\varepsilon} \right),$$

i. e. $k = O \left(\frac{1}{\sqrt{\varepsilon}} \ln \left(\frac{1}{\varepsilon} \right) \right)$ iterations.

In analogy to the considerations made in Section 4.3 we obtain for all $k \geq 0$

$$\begin{aligned} \|\nabla\theta(p_k)\| &\leq 2 \sqrt{L(\kappa)(\theta(0) - \theta(p^*))} e^{-\frac{k}{2} \sqrt{\frac{\kappa}{L(\kappa)}}} + 2\kappa R \\ &= 2 \sqrt{L(\kappa)(\theta(0) - \theta(p^*))} e^{-\frac{k}{2} \sqrt{\frac{\kappa}{L(\kappa)}}} + \frac{2\varepsilon}{R}. \end{aligned}$$

Therefore, in order to guarantee $\|\nabla\theta(p_k)\| \leq \frac{3\varepsilon}{R}$, we need $k = O \left(\frac{1}{\sqrt{\varepsilon}} \ln \left(\frac{1}{\varepsilon} \right) \right)$ iterations, which coincide with the convergence rate for the dual objective values.

4.4.2 The case g is everywhere differentiable with Lipschitz continuous gradient

Assuming additionally to the standing assumptions that the function $g : \mathbb{R}^m \rightarrow \mathbb{R}$ has full domain and it is differentiable with $\frac{1}{\kappa}$ -Lipschitz continuous gradient, for $\kappa > 0$, the second smoothing, as done in Subsection 3.2 can be omitted. The fast gradient method (8) can be applied to the function $\theta_\rho : \mathbb{R}^m \rightarrow \mathbb{R}$, $\theta_\rho := f_\rho^*(A^*p) + g^*(-p)$, which is κ -strongly convex due to [1, Theorem 18.15] and differentiable with Lipschitz continuous gradient. The Lipschitz constant of $\nabla\theta_\rho$ is $L(\rho) := \frac{\|A\|^2}{\rho} + \frac{1}{\mu}$.

The algorithm (8) applied to θ_ρ states

$$\begin{aligned}\theta_\rho(p_k) - \theta_\rho(p_{DS}^*) &\leq \left(\theta_\rho(0) - \theta_\rho(p_{DS}^*) + \frac{\kappa}{2} \|0 - p_{DS}^*\|^2 \right) e^{-k\sqrt{\frac{\kappa}{L(\rho)}}} \\ &\leq 2(\theta_\rho(0) - \theta_\rho(p_{DS}^*)) e^{-k\sqrt{\frac{\kappa}{L(\rho)}}} \quad \forall k \geq 0.\end{aligned}$$

Since $\theta_\rho(0) \stackrel{(5)}{\leq} \theta(0)$ and $\theta_\rho(p_{DS}^*) \stackrel{(5)}{\geq} \theta(p_{DS}^*) - \rho D_f \geq \theta(p^*) - \rho D_f$, we obtain

$$\theta_\rho(0) - \theta_\rho(p_{DS}^*) \leq \theta(0) - \theta(p^*) + \rho D_f. \quad (29)$$

On the other hand, since $\theta_\rho(p_k) - \theta_\rho(p_{DS}^*) \stackrel{(5)}{\geq} \theta(p_k) - \rho D_f - \theta(p^*)$, it follows

$$\begin{aligned}\theta(p_k) - \theta(p^*) &\leq \rho D_f + \theta_\rho(p_k) - \theta_\rho(p_{DS}^*) \\ &\leq \rho D_f + 2(\theta(0) - \theta(p^*) + \rho D_f) e^{-k\sqrt{\frac{\kappa}{L(\rho)}}} \quad \forall k \geq 0.\end{aligned}$$

Hence, when $\varepsilon > 0$, in order to guarantee ε -optimality for the dual objective, we force both terms in the above estimate less than or equal to $\frac{\varepsilon}{2}$. By taking

$$\rho := \rho(\varepsilon) = \frac{\varepsilon}{2D_f}, \quad (30)$$

in contrast to (23), we need

$$k \geq \sqrt{\frac{L(\rho)}{\kappa}} \ln \left(\frac{4(\theta(0) - \theta(p^*) + \frac{\varepsilon}{2})}{\varepsilon} \right),$$

i. e. $k = O\left(\frac{1}{\sqrt{\varepsilon}} \ln\left(\frac{1}{\varepsilon}\right)\right)$ iterations.

We obtain as well

$$\begin{aligned}\|\nabla\theta_\rho(p_k)\| &\stackrel{(11)}{\leq} 2\sqrt{L(\rho)(\theta_\rho(0) - \theta_\rho(p_{DS}^*))} e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho)}}} \\ &\stackrel{(29)}{\leq} 2\sqrt{L(\rho)(\theta(0) - \theta(p^*) + \rho D_f)} e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho)}}} \\ &\stackrel{(30)}{=} 2\sqrt{L(\rho)(\theta(0) - \theta(p^*) + \frac{\varepsilon}{2})} e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho)}}} \quad \forall k \geq 0.\end{aligned}$$

Therefore, in order to guarantee $\|\nabla\theta(p_k)\| \leq \frac{3\varepsilon}{R}$, we need $k = O\left(\frac{1}{\sqrt{\varepsilon}} \ln\left(\frac{1}{\varepsilon}\right)\right)$ iterations, which is the same convergence rate as for the dual objective values.

4.4.3 The case f is strongly convex and g is everywhere differentiable with Lipschitz continuous gradient

Assuming additionally to the standing assumptions that the function $f : \mathcal{H} \rightarrow \overline{\mathbb{R}}$ is ρ -strongly convex, for $\rho > 0$, and the function $g : \mathbb{R}^m \rightarrow \mathbb{R}$ has full domain and it is differentiable with $\frac{1}{\kappa}$ -Lipschitz continuous gradient, for $\kappa > 0$, both the first and second smoothing can be omitted. The fast gradient method (8) can be applied to the function

$\theta : \mathbb{R}^m \rightarrow \mathbb{R}$, $\theta := f^*(A^*p) + g^*(-p)$, which is κ -strongly convex and differentiable with Lipschitz continuous gradient. The Lipschitz constant of $\nabla\theta$ is $L := \frac{\|A\|^2}{\rho} + \frac{1}{\mu}$.

The fast gradient scheme (8) applied to θ yields for all $k \geq 0$

$$\theta(p_k) - \theta(p^*) \stackrel{(9)}{\leq} (\theta(0) - \theta(p^*) + \frac{\kappa}{2} \|0 - p^*\|^2) e^{-k\sqrt{\frac{\kappa}{L}}} \stackrel{(10)}{\leq} 2(\theta(0) - \theta(p^*)) e^{-k\sqrt{\frac{\kappa}{L}}}$$

and, from here, when $\varepsilon > 0$,

$$2(\theta(0) - \theta(p^*)) e^{-k\sqrt{\frac{\kappa}{L}}} \leq \varepsilon \Leftrightarrow k \geq \sqrt{\frac{L}{\kappa}} \ln \left(\frac{2(\theta(0) - \theta(p^*))}{\varepsilon} \right)$$

On the other hand, formula (11) states $\|\nabla\theta(p_k)\| \leq 2\sqrt{L(\theta(0) - \theta(p^*))} e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L}}}$ for all $k \geq 0$, thus

$$2\sqrt{L(\theta(0) - \theta(p^*))} e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L}}} \leq \varepsilon \Leftrightarrow k \geq 2\sqrt{\frac{L}{\kappa}} \ln \left(\frac{2\sqrt{L(\theta(0) - \theta(p^*))}}{\varepsilon} \right).$$

In conclusion, in order to guarantee ε -accuracy for the dual objective values and for the decrease of $\|\nabla\theta(\cdot)\|$ to 0, we need $O\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$ iterations.

4.5 Constructing an approximate primal solution

In the remaining of this section we work in the setting of our initial standing assumptions and show, first of all, how to recover approximately optimal solutions for the primal (P) from the sequence of approximately dual solutions $(p_k)_{k \geq 0}$. This will be followed by a convergence analysis for the approximate primal optimal solutions. One can easily notice that the investigations made here remain valuable when working in the special settings of the previous section, too.

Since our main focus is to solve the primal optimization problem (P) , we prove as follows that the sequences $(x_{f,p_k})_{k \geq 0} \subseteq \text{dom } f$ and $(x_{g,p_k})_{k \geq 0} \subseteq \text{dom } g$ constructed in Subsection 3.1 contain all the information one needs to recover approximately optimal solutions to (P) .

Since $\theta_\rho(p_k) - \theta(p^*) \stackrel{(5)}{\leq} \theta(p_k) - \theta(p^*) \leq \varepsilon$ and

$$\theta_\rho(p_k) - \theta(p^*) \stackrel{(5)}{\geq} \theta(p_k) - \rho D_f - \theta(p^*) \stackrel{(22)}{=} \underbrace{\theta(p_k) - \theta(p^*)}_{\geq 0} - \frac{\varepsilon}{3} \geq -\frac{\varepsilon}{3},$$

it holds $|\theta_\rho(p_k) - \theta(p^*)| \leq \varepsilon$ for all $k \geq 0$. Further, for $p_k \in \mathbb{R}^m$ we have

$$\begin{aligned} \theta_\rho(p_k) &= f_\rho^*(A^*p_k) + g^*(-p_k) \\ &= \langle p_k, Ax_{f,p_k} \rangle - f(x_{f,p_k}) - \frac{\rho}{2} \|x_{f,p_k}\|^2 - \langle p_k, x_{g,p_k} \rangle - g(x_{g,p_k}) \end{aligned}$$

and from here (notice that $-v(D) = \theta(p^*)$)

$$f(x_{f,p_k}) + g(x_{g,p_k}) - v(D) = \langle p_k, \nabla\theta_\rho(p_k) \rangle + (\theta(p^*) - \theta_\rho(p_k)) - \frac{\rho}{2} \|x_{f,p_k}\|^2 \quad \forall k \geq 0.$$

It follows

$$\begin{aligned}
|f(x_{f,p_k}) + g(x_{g,p_k}) - v(D)| &\leq \|p_k\| \|\nabla \theta_\rho(p_k)\| + |\theta(p^*) - \theta_\rho(p_k)| + \frac{\rho}{2} \|x_{f,p_k}\|^2 \\
&\leq \|p_k\| \|\nabla \theta_\rho(p_k)\| + \epsilon + \rho D_f \\
&\stackrel{(22)}{\leq} \|p_k\| \|\nabla \theta_\rho(p_k)\| + 2\epsilon \stackrel{(28)}{\leq} \frac{2\epsilon}{R} \|p_k\| + 2\epsilon \quad \forall k \geq 0.
\end{aligned}$$

Further, $\|p_k\|$ can be estimated above using

$$\|p_k\| = \|p_k + p_{DS}^* - p_{DS}^*\| \leq \|p_k - p_{DS}^*\| + \|p_{DS}^*\| \stackrel{(13)}{\leq} 2 \|p_{DS}^*\| \stackrel{(26)}{\leq} 2\sqrt{2}R,$$

therefore, we obtain

$$|f(x_{f,p_k}) + g(x_{g,p_k}) - v(D)| \leq 4\sqrt{2}\epsilon + 2\epsilon = 2(2\sqrt{2} + 1)\epsilon \quad \forall k \geq 0. \quad (31)$$

By taking into account weak duality, i.e. $v(D) \leq v(P)$, we conclude that $x_{f,p_k} \in \text{dom } f$ and $x_{g,p_k} \in \text{dom } g$ can be seen as approximately optimal solutions to (P) when k is high enough to satisfy (28).

4.6 Existence of an optimal solution

This section is devoted to the convergence analysis of our primal sequences when ϵ converges to zero. To this end let $(\epsilon_n)_{n \geq 0} \subseteq \mathbb{R}_+$ be a decreasing sequence of positive scalars with $\lim_{n \rightarrow \infty} \epsilon_n = 0$. For each $n \geq 0$, the double smoothing algorithm (8) with smoothing parameters ρ_{ϵ_n} and κ_{ϵ_n} given by (22) requires at least $k = k(\epsilon_n)$ iterations to fulfill (28). For $n \geq 0$ we denote

$$\bar{x}_n := x_{f,p_{k(\epsilon_n)}} \in \text{dom } f \text{ and } \bar{y}_n := x_{g,p_{k(\epsilon_n)}} \in \text{dom } g.$$

Due to the boundedness of $\text{dom } f$, its closure $\text{cl}(\text{dom } f)$ is weakly compact (see [1, Theorem 3.3]) and there exists a subsequence $(\bar{x}_{n_l})_{l \geq 0}$ and $\bar{x} \in \mathcal{H}$ such that \bar{x}_{n_l} weakly converges to $\bar{x} \in \text{cl}(\text{dom } f)$ when $l \rightarrow +\infty$. Since $A : \mathcal{H} \rightarrow \mathbb{R}^m$ is linear and continuous, the sequence $A\bar{x}_{n_l}$ will converge to $A\bar{x}$ when $l \rightarrow +\infty$. In view of relation (28) we get

$$0 \leq \|A\bar{x}_{n_l} - \bar{y}_{n_l}\| \leq \frac{2\epsilon_{n_l}}{R} \quad \forall l \geq 0. \quad (32)$$

This means that the sequence $(\bar{y}_{n_l})_{l \geq 0} \subseteq \text{dom } g$ is obviously bounded, hence there exists a subsequence of it (still denoted by $(\bar{y}_{n_l})_{l \geq 0}$) and an element $\bar{y} \in \text{cl}(\text{dom } g)$ such that $\bar{y}_{n_l} \rightarrow \bar{y}$ when $l \rightarrow +\infty$. Taking $l \rightarrow +\infty$ in (32) it follows $A\bar{x} = \bar{y}$. Furthermore, due to (31), we have

$$f(\bar{x}_{n_l}) + g(\bar{y}_{n_l}) \leq v(D) + 2(3\sqrt{2} + 1)\epsilon_{n_l} \quad \forall l \geq 0$$

and, by using the lower semicontinuity of f and g and [1, Theorem 9.1], we obtain

$$\begin{aligned}
f(\bar{x}) + g(A\bar{x}) &\leq \liminf_{l \rightarrow \infty} \{f(\bar{x}_{n_l}) + g(\bar{y}_{n_l})\} \\
&\leq \lim_{l \rightarrow \infty} \{v(D) + 2(3\sqrt{2} + 1)\epsilon_{n_l}\} = v(D) \leq v(P).
\end{aligned}$$

Since $v(P) \in \mathbb{R}$, we have $\bar{x} \in \text{dom } f$ and $A\bar{x} \in \text{dom } g$, which yields that \bar{x} is an optimal solution to (P) .

5 Two examples in image processing

In this section we are solving a linear inverse problem which arises in the field of signal and image processing via the double smoothing algorithm developed in this paper. For a given matrix $A \in \mathbb{R}^{n \times n}$ describing a *blur operator* and a given vector $b \in \mathbb{R}^n$ representing the *blurred and noisy image* the task is to estimate the *unknown original image* $x^* \in \mathbb{R}^n$ fulfilling

$$Ax = b.$$

To this end we make use of two regularization functionals with different properties.

5.1 An l_1 regularization problem

We start by solving the l_1 regularized convex optimization problem

$$(P) \quad \inf_{x \in S} \left\{ \|Ax - b\|^2 + \lambda \|x\|_1 \right\},$$

where $S \subseteq \mathbb{R}^n$ is an n -dimensional cube representing the range of the pixels and $\lambda > 0$ the regularization parameter. The problem to be solved can be equivalently written as

$$(P) \quad \inf_{x \in \mathbb{R}^n} \{f(x) + g(Ax)\},$$

for $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, $f(x) = \lambda \|x\|_1 + \delta_S(x)$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}$, $g(y) = \|y - b\|^2$. Thus f is proper, convex and lower semicontinuous with bounded domain and g is a 2-strongly convex function with full domain, differentiable everywhere and with Lipschitz continuous gradient having as Lipschitz constant 2. This means that we are in the setting of Subsection 4.4.2.

By making use of gradient methods, both the iterative shrinkage-thresholding algorithm (ISTA) (see [9]) and its accelerated variant FISTA (see [2, 3]) solve the optimization problem (P) in $O\left(\frac{1}{\epsilon}\right)$ and $O\left(\frac{1}{\sqrt{\epsilon}}\right)$ iterations, respectively, whereas the convergence rate of our method is $O\left(\frac{1}{\sqrt{\epsilon}} \ln\left(\frac{1}{\epsilon}\right)\right)$.

Since each pixel furnishes a greyscale value which is between 0 and 255, a natural choice for the convex set S would be the n -dimensional cube $[0, 255]^n \subseteq \mathbb{R}^n$. In order to reduce the Lipschitz constant which appears in the developed approach, we scale the pictures to which refer within this subsection such that each of their pixels ranges in the interval $\left[0, \frac{1}{10}\right]$. We concretely look at the 256×256 *cameraman test image*, which is part of the image processing toolbox in Matlab. The dimension of the vectorized and scaled cameraman test image is $n = 256^2 = 65536$. By making use of the Matlab functions `imfilter` and `fspecial`, this image is blurred as follows:

```

1 H=fspecial('gaussian',9,4);    % gaussian blur of size 9 times 9
2                               % and standard deviation 4
3 B=imfilter(X,H,'conv','symmetric'); % B=observed blurred image
4                               % X=original image

```

In row 1 the function `fspecial` returns a rotationally symmetric Gaussian lowpass filter of size 9×9 with standard deviation 4. The entries of H are nonnegative and their sum

adds up to 1. In row 3 the function `imfilter` convolves the filter H with the image $X \in \mathbb{R}^{256 \times 256}$ and outputs the blurred image $B \in \mathbb{R}^{256 \times 256}$. The boundary option "symmetric" corresponds to reflexive boundary conditions.

Thanks to the rotationally symmetric filter H , the linear operator $A \in \mathbb{R}^{n \times n}$ given by the Matlab function `imfilter` is symmetric, too. By making use of the real spectral decomposition of A , it shows that $\|A\|^2 = 1$. After adding a zero-mean white Gaussian noise with standard deviation 10^{-4} , we obtain the blurred and noisy image $b \in \mathbb{R}^n$ which is shown in Figure 5.1.



Figure 5.1: The 256×256 cameraman test image

The dual optimization problem in minimization form is

$$(D) \quad - \inf_{p \in \mathbb{R}^n} \{f^*(A^*p) + g^*(-p)\}$$

and, due to the fact that g has full domain, strong duality for (P) and (D) holds, i.e. $v(P) = v(D)$ and (D) has an optimal solution (see, for instance, [5, 6]). By taking into consideration (30), the smoothing parameter is taken as

$$\rho := \frac{\epsilon}{2D_f} \tag{33}$$

for $D_f = \sup \left\{ \frac{\|x\|^2}{2} : x \in \left[0, \frac{1}{10}\right]^n \right\} = 327.68$, while the accuracy is chosen to be $\epsilon = 0.3$ and the regularization parameter is set to $\lambda = 2e-6$.

We show next that the sequences of approximate primal solutions $(x_{f,p_k})_{k \geq 0}$ and $(x_{g,p_k})_{k \geq 0}$ can be easily calculated. Indeed, for $k \geq 0$ we have

$$\begin{aligned} x_{f,p_k} &= \arg \min_{x \in [0, \frac{1}{10}]^n} \left\{ \lambda \|x\|_1 + \frac{\rho}{2} \left\| \frac{A^*p_k}{\rho} - x \right\|^2 \right\} \\ &= \arg \min_{x \in [0, \frac{1}{10}]^n} \left\{ \sum_{i=1}^n \left[\lambda |x_i| + \frac{\rho}{2} \left(\frac{(A^*p_k)_i}{\rho} - x_i \right)^2 \right] \right\} \end{aligned}$$

and, in order to determine it, we need to solve the one-dimensional convex optimization

problem

$$\inf_{x_i \in [0, \frac{1}{10}]} \left\{ \lambda x_i + \frac{\rho}{2} \left(\frac{(A^* p_k)_i}{\rho} - x_i \right)^2 \right\},$$

for $i = 1, \dots, n$, which has as unique optimal solution $\mathcal{P}_{[0, \frac{1}{10}]} \left(\frac{1}{\rho} ((A^* p_k)_i - \lambda) \right)$. Thus,

$$x_{f,p_k} = \mathcal{P}_{[0, \frac{1}{10}]^n} \left(\frac{1}{\rho} (A^* p_k - \lambda \mathbb{1}^n) \right).$$

On the other hand, for all $k \geq 0$ we have

$$x_{g,p_k} = \arg \min_{x \in \mathbb{R}^n} \{ \langle p_k, x \rangle + g(x) \} = \arg \min_{x \in \mathbb{R}^n} \{ \langle p_k, x \rangle + \|x - b\|^2 \} = b - \frac{1}{2} p_k.$$



Figure 5.2: Iterations of ISTA, FISTA and double smoothing (DS) for solving (P)

Figure 5.2 shows the iterations 50 and 100 of ISTA, FISTA and the double smoothing (DS) approach. The objective function values at iteration k are denoted by ISTA_k , FISTA_k and, respectively, DS_k (e.g. $\text{DS}_k := f(x_{f,p_k}) + g(Ax_{f,p_k})$). All in all, the visual quality of the restored cameraman image after 100 iterations, when using FISTA or DS, is quite comparable, whereas the recovered image by ISTA is still blurry. However, a valuable tool for measuring the quality of these images is the so-called *improvement in*

signal-to-noise ratio (ISNR), which is defined as

$$\text{ISNR}(k) = 10 \log_{10} \left(\frac{\|x - b\|^2}{\|x - x_k\|^2} \right)$$

where x , b and x_k denote the original, observed and estimated image at iteration k , respectively. Figure 5.3 shows the evolution of the ISNR values when using DS, FISTA and ISTA to solve (P).

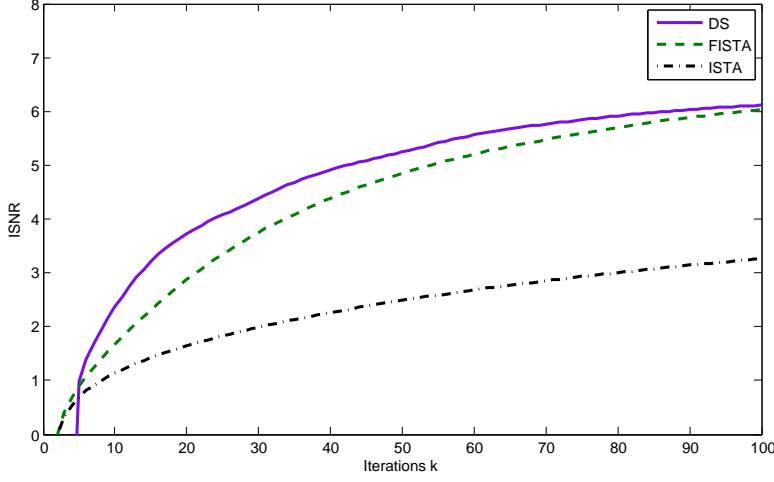


Figure 5.3: Improvement in signal-to-noise ratio (ISNR)

5.2 An $l_2 - l_1$ regularization problem

The second convex optimization problem we solve is

$$(P) \quad \inf_{x \in S} \left\{ \|Ax - b\|^2 + \lambda(\|x\|^2 + \|x\|_1) \right\},$$

where $S \subseteq \mathbb{R}^n$ is the n -dimensional cube $[0, 1]^n$ representing the pixel range, $\lambda > 0$ the regularization parameter and $\|\cdot\|^2 + \|\cdot\|_1$ the regularization functional, already used in [7]. The problem to be solved can be equivalently written as

$$(P) \quad \inf_{x \in \mathbb{R}^n} \{f(x) + g(Ax)\},$$

for $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, $f(x) = \lambda(\|x\|^2 + \|x\|_1) + \delta_S(x)$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}$, $g(y) = \|y - b\|^2$. Thus f is proper, 2λ -strongly convex and lower semicontinuous with bounded domain and g is a 2-strongly convex function with full domain, differentiable everywhere and with Lipschitz continuous gradient having as Lipschitz constant 2. This time we are in the setting of the Subsection 4.4.3, the Lipschitz constant of the gradient of $\theta : \mathbb{R}^n \rightarrow \mathbb{R}$, $\theta(p) = f^*(A^*p) + g^*(-p)$, being $L = \frac{1}{2\lambda} + \frac{1}{2}$. By applying the double smoothing approach one obtains a rate of convergence of $O\left(\ln\left(\frac{1}{\epsilon}\right)\right)$ for solving (P).

In this example we take a look at the *blobs test image* shown in Figure 5.4 which is also part of the image processing toolbox in Matlab. The picture undergoes the

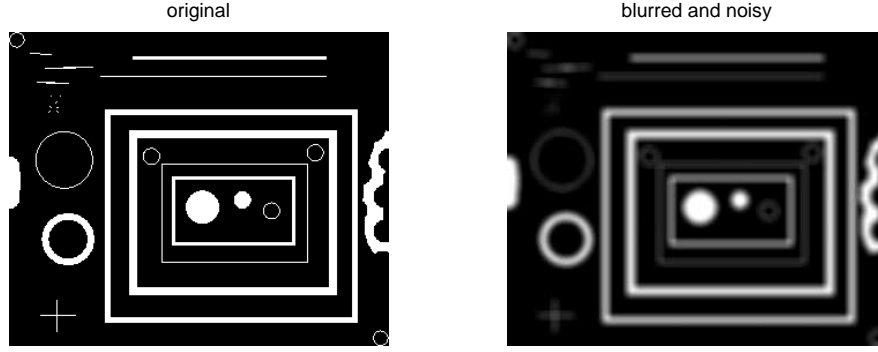


Figure 5.4: The 272×329 blobs test image

same blur as described in the previous section. Since our pixel range has changed, we now use additive zero-mean white Gaussian noise with standard deviation 10^{-3} and the regularization parameter is changed to $\lambda = 2e-5$.

We calculate next the sequences of approximate primal solutions $(x_{f,p_k})_{k \geq 0}$ and $(x_{g,p_k})_{k \geq 0}$. Indeed, for $k \geq 0$ we have

$$\begin{aligned} x_{f,p_k} &= \arg \min_{x \in [0,1]^n} \left\{ \lambda \|x\|^2 + \lambda \|x\|_1 - \langle A^* p_k, x \rangle \right\} \\ &= \arg \min_{\substack{i=1,\dots,n \\ x_i \in [0,1]}} \left\{ \sum_{i=1}^n \left[-(A^* p_k)_i x_i + \lambda x_i^2 + \lambda x_i \right] \right\} = \mathcal{P}_{[0,1]^n} \left(\frac{1}{2\lambda} (A^* p_k - \lambda \mathbb{1}^n) \right). \end{aligned}$$

and

$$x_{g,p_k} = \arg \min_{x \in \mathbb{R}^n} \{ \langle p_k, x \rangle + g(x) \} = \arg \min_{x \in \mathbb{R}^n} \left\{ \langle p_k, x \rangle + \|x - b\|^2 \right\} = b - \frac{1}{2} p_k.$$

Figure 5.5 shows the iterations 50 and 100 of ISTA, FISTA and the double smoothing (DS) technique together with the corresponding function values denoted by ISTA_k , FISTA_k or DS_k . As before, the function values of FISTA are slightly lower than those of DS, while ISTA is far behind these methods, not only from theoretical point of view, but also as it can be detected visually. Figure 5.6 displays the improvement in signal-to-noise ration for ISTA, FISTA and DS and it shows that DS outperforms the other two methods from the point of view of the quality of the reconstruction.

6 Conclusions

In this article we investigate the possibilities of accelerating the double smoothing technique when solving unconstrained nondifferentiable convex optimization problems. This method, which assumes the minimization of the doubly regularized Fenchel dual objective, allows in the most general case to reconstruct an approximately optimal primal solution in $O\left(\frac{1}{\epsilon} \ln\left(\frac{1}{\epsilon}\right)\right)$ iterations. We show that under some appropriate assumptions for the functions involved in the formulation of the problem to be solved this convergence rate can be improved to $O\left(\frac{1}{\sqrt{\epsilon}} \ln\left(\frac{1}{\epsilon}\right)\right)$, or even to $O\left(\ln\left(\frac{1}{\epsilon}\right)\right)$.

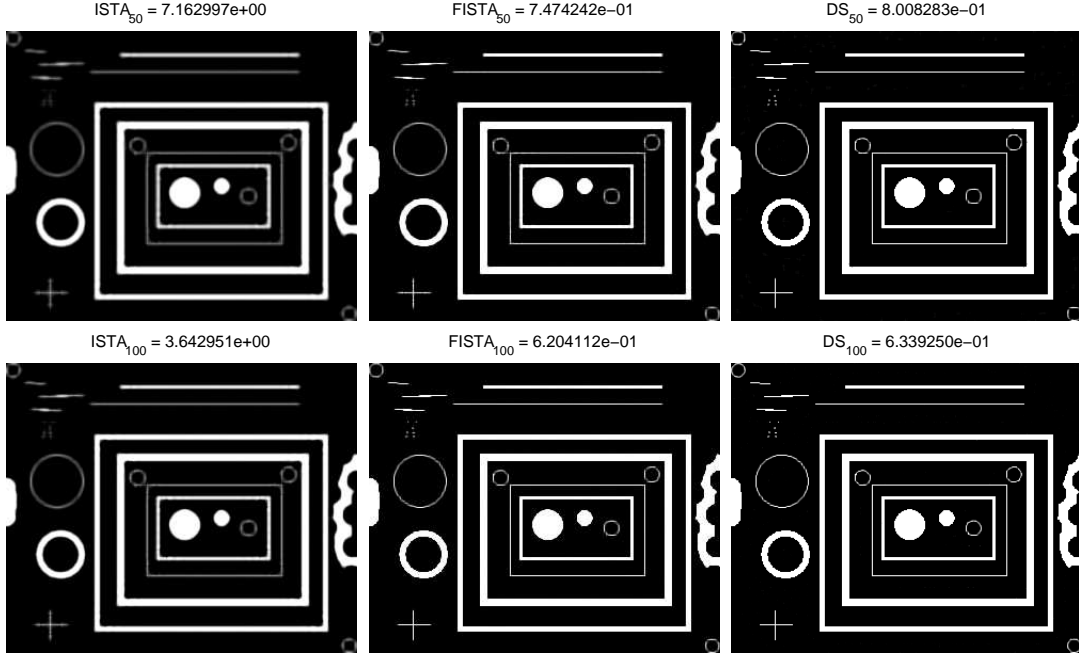


Figure 5.5: Iterations of ISTA, FISTA and double smoothing (DS) for solving (P)

References

- [1] H.H. Bauschke and P.L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics, Springer, 2011.
- [2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- [3] A. Beck and M. Teboulle. Gradient-based algorithms with applications to signal recovery problems. In: Y. Eldar and D. Palomar (eds.), “Convex Optimization in Signal Processing and Communications”, pp. 33–88. Cambridge University Press, 2010.
- [4] J.F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research and Financial Engineering, 2000.
- [5] R.I. Boş. *Conjugate Duality in Convex Optimization*. Lecture Notes in Economics and Mathematical Systems, Vol. 637, Springer-Verlag Berlin Heidelberg, 2010.
- [6] R.I. Boş, S.-M. Grad and G. Wanka. *Duality in Vector Optimization*. Springer-Verlag Berlin Heidelberg, 2009.
- [7] R.I. Boş and T. Hein. Iterative regularization with general penalty term theory and application to L^1 - and TV -regularization. to appear in *Inverse Problems*, 2012.
- [8] R.I. Boş and C. Hendrich. A double smoothing technique for solving unconstrained nondifferentiable convex optimization problems. *arXiv:1203.2070v1 [math.OC]*, 2012.

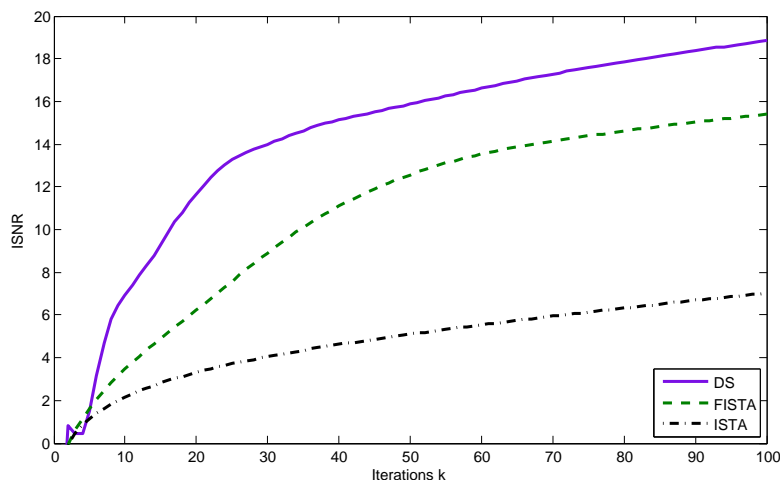


Figure 5.6: Improvement in signal-to-noise ratio (ISNR)

- [9] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.
- [10] O. Devolder, F. Glineur and Y. Nesterov. A double smoothing technique for constrained convex optimization problems and applications to optimal control. *Optimization Online*, http://www.optimization-online.org/DB_FILE/2011/01/2896.pdf, 2010.
- [11] O. Devolder, F. Glineur and Y. Nesterov. Double smoothing technique for infinite-dimensional optimization problems with applications to optimal control. *CORE Discussion Paper*, http://www.uclouvain.be/cps/ucl/doc/core/documents/coredp2010_34web.pdf, 2010.
- [12] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.
- [13] Y. Nesterov. Excessive gap technique in nonsmooth convex optimization. *SIAM Journal of Optimization*, 16(1):235–249, 2005.
- [14] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.
- [15] Y. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming*, 110(2):245–259, 2005.